Predicting car accident severity

Cauê de Ulhôa Castro

September 2020

1. Introduction

1.1. The problem

Too many traffic accidents have been reported in the road system of Seattle.

1.2. Solution

The city's department of transportation, along with the budget office, would be the main target audience to be convinced that the project worth to be invested on.

Enhancing awareness is proven to be directly correlated with decrease in the number of accidents. However, some types of awareness messages are more effective than others. Usually, general statements – such as: "Fast driving kills" – are not as effective as when real statistics is attached to it – like: "Driving above speed limit causes 50% more severe accidents – where the person reading it can make a better sense of the risks involved though quantitative information.

With that in mind, the idea is to gather as many data as possible on traffic accidents in the roads of Seattle and group it by different degrees of severity (e.g. fatality, serious injury, injury, prop damage). Then, predict the probability of a road user to get involved in those accidents given different conditions. For instance:

Condition	Fatality	Serious injury	Injury	Prop damage
Raining				
Sunny				
Peak hours				
Weekends				
Above speed limit				
Drunk				

These predicted probabilities would then be displayed on road signs according with the current weather condition (for instance) to warn drivers of the potential dangers lying ahead, prompting safe driving practices.

A baseline should be established in order to assess how much higher would be the chances of accident when the conditions are changed. Before exploring the dataset is already possible to predict that the baseline should be done under good weather conditions, no alcohol or drugs consumption, no peak hour, weekdays and not above speed limit.

2. The dataset

The dataset to be used in this project consists of almost 200.000 accident records and 38 attributes. The attribute to be predicted is named SEVERITYCODE, consisting of different code numbers referring to its respective level of severity, separated as Fatality (1), Serious injury (2a), Injury (2b) and Property Damage (3). The timeframe of the data set is from 2004 to present.

The dataset was download from the Department of Transportation of the city of Seattle, under ArcGIS Metadata Form.

Next, we discuss about the attributes relevant to reach the objective.

'COLLISIONTYPE' describes the type of collision.

```
Parked Car    47987
Angles        34674
Rear Ended    34090
Other         23703
Sideswipe     18609
Left Turn     13703
Pedestrian     6608
Cycles         5415
Right Turn     2956
Head On        2024
```

'VEHCOUNT' is the number of vehicles involved in the collision. It might have a correlation with severity.

'INCDATE' the date of the incident. It is worth investigating if different days of the week or holidays play a significant role on the number of incidents.

'INCDTTM' date and time of the incident. Time also might play a role. Rush hour and evening might cause more accidents? LIGHTCOND describes the light conditions at the time, so it can also be used for this analysis. Choose one of them.

'JUNCTIONTYPE' worth investigating correlation with severity

Mid-Block (not related to intersection)	89800
At Intersection (intersection related)	62810
Mid-Block (but intersection related)	22790
Driveway Junction	10671
At Intersection (but not related to intersection)	2098
Ramp Junction	166
Unknown	9

The attributes below refer to external conditions, not influenced by the driver. Accumulative statistics must be conducted for the case when two or more conditions are met.

WEATHER A description of the weather conditions during the time of the collision.

ROADCOND The condition of the road during the collision.

LIGHTCOND The light conditions during the collision.

The attributes below refer to the driver’s that might have potentially influenced in the accident.

SPEEDING Whether or not speeding was a factor in the collision. (Y/N) – Won’t go in because the majority is NaN

‘INATTENTIONIND’ Whether or not collision was due to inattention (Y/N). It might be correlated with severity.

‘UNDERINFL’ Whether or not a driver involved was under the influence of drugs or alcohol. This variable deserves an analysis on its own, in order to highlight statistics for these specific cases.

HITPARKEDCAR. Whether or not the collision involved hitting a parked car (Y/N). Should these cases be excluded from the list?

3. Methodology

Since all attributes selected are categorical, with the exception of the ones referring to the number of vehicles and people involved in the accident, a sum of both ‘yes’ and ‘no’ dummy variables were used to explore each feature in relation with the target. A second filter in the data was performed in this step, where features with too little count (less than 5% of the total) were deleted because it presented no clear correlation with the severity and could cause noise in the machine learning results. The remaining features were analyzed focusing on its correlation with the target result.

SEVERITYCODE	**1.0**	**2.0**	**% severe**
UNDERINFL	5477.0	3533.0	1.550241
VEHCOUNT	252536.0	107788.0	2.342895
PERSONCOUNT	304777.0	156768.0	1.944128
sandy	37.0	12.0	3.083333
Clear	88475.0	37450.0	2.362483
fog	369.0	187.0	1.973262
Overcast	18530.0	8681.0	2.134547
cloudy	2.0	3.0	0.666667
Raining	21601.0	11097.0	1.946562
crosswind	18.0	7.0	2.571429

sleet	82.0	27.0	3.037037
Snowing	719.0	168.0	4.279762
road_dry	96856.0	41488.0	2.334555
road_ice	900.0	263.0	3.422053
road_oil	35.0	24.0	1.458333
road_sand	45.0	22.0	2.045455
road_snow	811.0	166.0	4.885542
road_pond	79.0	30.0	2.633333
road_wet	31107.0	15639.0	1.989066
DAYTIME	77357.0	37451.0	2.065552
NIGHTTIME	52476.0	20181.0	2.600268
at_intersect_n	1469.0	621.0	2.365539
at_intersect	35544.0	27121.0	1.310571
Driveway	7395.0	3228.0	2.290892
midblock	15438.0	7280.0	2.120604
midblock_n	69876.0	19329.0	3.615086
ramp	111.0	53.0	2.094340

Here, the expected correlations were verified in order to double check whether a clear correlation could be verified or not. If no influence in the severity were identified, the variable would also have to get discarded.

After the exploratory analysis were completed, the features regarded as useful to the model were:

- SEVERITYCODE
- UNDERINFL – whether or not the driver was under influence of alcohol
- VEHCOUNT – number of vehicles involved in the car accident
- PERSONCOUNT – number of people involved in the car accident
- TIME OF THE DAY – whether the accident took place during the day or night. This feature was set by the author according with the time data
- WEATHER – weather condition
- ROADCOND – road condition
- JUNCTIONTYPE – at intersection or midblock

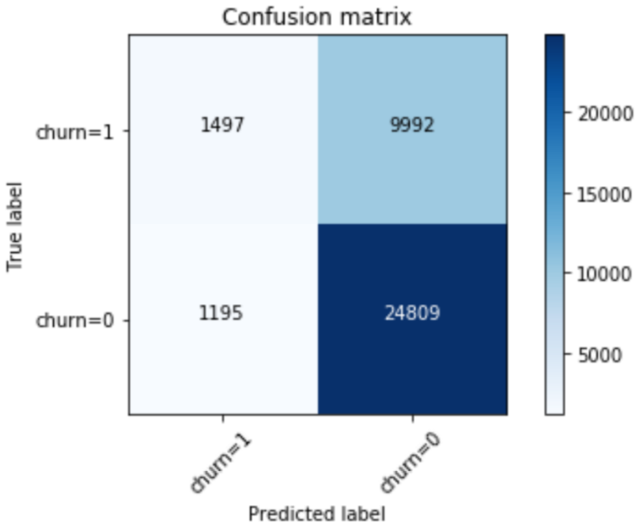
Then, the machine learning technique selected was the Logistic Regression. This is a classification supervised model. The reason behind this choice is because it is a widely used model to make predictions to categorical target label and, furthermore, calculates the probability of a certain category to happen. This last result essential for the objective of this study, since we want to know the probability of an accident to happen for different scenarios.

Three different evaluation algorithms were used in order to assess the accuracy of the results obtained by the Logistic Regression model: Jaccard Index, F1 Score and Log Loss.

Jaccard index returned a value of 0.70 and the Log Loss metric, 0.58. The table below summarizes the results obtained by the F1 score method.

	precision	recall	f1-score	support
0	0.71	0.95	0.82	26004
1	0.56	0.13	0.21	11489
accuracy			0.70	37493
macro avg	0.63	0.54	0.51	37493
weighted avg	0.66	0.70	0.63	37493

Also, a confusion matrix was computed to visualize the evaluation of false negatives and false positives.



4. Results

The model was used to predict many different scenarios. The table below summarizes the results for each of those hypothetic scenarios in order to assess increase in severity probability with increase in exposure to more extreme conditions.

WEATHER	ROADCOND	DAYTIME	UNDERINFL	VEHCOUNT	PERSONCOUNT	JUNCTIONTYPE	SEVERITY	PROBAB.
snowing	snow	night	no	2	2	no mid block	0	72%
snowing	snow	night	yes	2	2	no mid block	0	70%
snowing	snow	day	no	2	2	no mid block	0	69%
snowing	snow	day	yes	2	2	no mid block	0	66%
raining	wet	night	no	2	2	no mid block	0	71%
raining	wet	night	yes	2	2	no mid block	0	68%
raining	wet	day	no	2	2	no mid block	0	67%
raining	wet	day	yes	2	2	no mid block	0	64%
clear	dry	night	no	2	2	no mid block	0	72%
clear	dry	night	yes	2	2	no mid block	0	69%
clear	dry	day	no	2	2	no mid block	0	69%
clear	dry	day	yes	2	2	no mid block	0	66%

5. Discussion

Jaccard index has returned acceptable results, showing that around 70% of the values were correctly predicted. Log Loss, on the other hand, showed relatively low results of 58%. Log Loss tells us how likely the model thinks the actually observed set of outcomes was, so it is a trickier metric to judge whether a low number indicates a good outcome or not.

The F1 score results show that the model performed well for the severity code 0 but not as good for the severity code 1, with low precision and recall results. Probably, maintaining the road condition along with weather condition might have influenced the poor results of the model, since both situations basically give the same information, causing doubled similar data outcomes.

When applying the model to hypothetic scenarios, all of those scenarios resulted in severity = 0, which lies in the less severe class. Considering that those scenarios range from less severe (e.g. dry road, clear weather, day time and no alcohol influence) to the most severe situations, it is clear that the model is not predicting severity accurately and probably some step was not performed properly.

6. Conclusion

Considering the evaluation of the model against the results obtained from the hypothetic scenarios, the model is currently not reliable enough to calculate the increase in probability of more severe accidents taking place when variables such as weather, time and alcohol influence, change.

Since the calculation involved mainly dummy variables, probably the model does not function as accurately as in cases where actual numbers are also involved in the model set up. However, since this particular situation was not discussed at length, it was not possible to neither spot nor fix the issue.